ORIE 4741: Learning with Big Messy Data

Exploratory Data Analysis

Professor Udell
Operations Research and Information Engineering
Cornell

September 8, 2020
Announcements

▶ If you’re taking lecture async: remember to submit participation post after each class!
▶ Sections start Wednesday. They are optional, attend any one you prefer. Section this week is a Julia tutorial.
▶ Office hours: Zoom links and times are posted on course website.
▶ Gradescope is open for submission of hw0, due Thursday 9-9-19 9:30am.
▶ First quiz this week! It should occupy about 20 minutes; you’ll have up to half an hour to complete it. Start it anytime between 6:15pm Thursday and 9pm Friday.

(All times ET)
Questions from campuswire

- search for your question before posting new question
- approximate date of voter registration is fine
Why Julia?

- the two language problem
- Julia is fast (JIT-compiled)
- Julia has pleasant syntax, esp for linear algebra (MATLAB-like, but more principled)
- Julia supports efficient parallelism (including multithreading)
- The Julia ecosystem

For this class: you can use any language you'd like (which your TAs can read), but the course staff will only support Julia.
Topics to review

We will cover (most of) these in section, too:

- **Linear algebra**: invertible matrices, rank, norm, basic matrix identities. When is a matrix invertible?
- **QR factorization**
- **Gradients (multivariate derivative)**
- **Projections**
- **SVD**
- **Maximum likelihood estimation**
- **Union bound**
- **Computational complexity**
Why look at the data?

- detect errors in data
- check assumptions
- select appropriate models
- understand relationships among the features
- understand relationships between features and labels
How to look at the data?

- inspect raw data
- summary statistics
- visualize
American community survey

2013 ACS:

- 3M respondents, 87 economic/demographic survey questions
  - income
  - cost of utilities (water, gas, electric)
  - weeks worked per year
  - hours worked per week
  - home ownership
  - looking for work
  - use foodstamps
  - education level
  - state of residence
  - ...

- 1/3 of responses missing

find it at https://people.orie.cornell.edu/mru8/orie4741/data/acs_2013.csv
How do computers work?

on a laptop:

- hard disk: usually $\leq 500$ GB
- memory (RAM): usually $\leq 16$ GB
- many programs (e.g., Excel): substantially more limited
How do computers work?

on a laptop:

- hard disk: usually $\leq 500$ GB
- memory (RAM): usually $\leq 16$ GB
- many programs (e.g., Excel): substantially more limited

don’t load a giant file into memory.
your computer will crash.
How do computers work?

on a laptop:

- hard disk: usually $\leq 500$ GB
- memory (RAM): usually $\leq 16$ GB
- many programs (e.g., Excel): substantially more limited

don’t load a giant file into memory.
your computer will crash.

how big is ACS data?
3M respondents $\times$ 100 questions $= 300$M numbers $\approx 300$MB
solution for large files: technology from the 70s!

bash shell:

▶ “how big are these files?”: ls -lh
▶ “show me some lines from the file”: head, tail, less
▶ “how many lines are in the file?”: wc -l
<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>HHTYPE</td>
<td>household type</td>
<td>categorical</td>
</tr>
<tr>
<td>STATEICP</td>
<td>state</td>
<td>categorical</td>
</tr>
<tr>
<td>OWNERSHP</td>
<td>own home</td>
<td>Boolean</td>
</tr>
<tr>
<td>COMMUSE</td>
<td>commercial use</td>
<td>Boolean</td>
</tr>
<tr>
<td>ACREHOUS</td>
<td>house on $\geq 10$ acres</td>
<td>Boolean</td>
</tr>
<tr>
<td>HHINCOME</td>
<td>household income</td>
<td>real</td>
</tr>
<tr>
<td>COSTELEC</td>
<td>monthly electricity bill</td>
<td>real</td>
</tr>
<tr>
<td>COSTWATR</td>
<td>monthly water bill</td>
<td>real</td>
</tr>
<tr>
<td>COSTGAS</td>
<td>monthly gas bill</td>
<td>real</td>
</tr>
<tr>
<td>FOODSTMP</td>
<td>on food stamps</td>
<td>Boolean</td>
</tr>
<tr>
<td>HCOVANY</td>
<td>have health insurance</td>
<td>Boolean</td>
</tr>
<tr>
<td>SCHOOL</td>
<td>currently in school</td>
<td>Boolean</td>
</tr>
<tr>
<td>EDUC</td>
<td>highest level of education</td>
<td>ordinal</td>
</tr>
<tr>
<td>GRADEATT</td>
<td>highest grade level attained</td>
<td>ordinal</td>
</tr>
<tr>
<td>EMPSTAT</td>
<td>employment status</td>
<td>categorical</td>
</tr>
<tr>
<td>LABFORCE</td>
<td>in labor force</td>
<td>Boolean</td>
</tr>
<tr>
<td>CLASSWKR</td>
<td>class of worker</td>
<td>Boolean</td>
</tr>
<tr>
<td>WKSWORK2</td>
<td>weeks worked per year</td>
<td>ordinal</td>
</tr>
<tr>
<td>UHRSWORK</td>
<td>usual hours worked per week</td>
<td>real</td>
</tr>
<tr>
<td>LOOKING</td>
<td>looking for work</td>
<td>Boolean</td>
</tr>
<tr>
<td>MIGRATE1</td>
<td>migration status</td>
<td>categorical</td>
</tr>
</tbody>
</table>
Julia and Jupyter

- Julia is a programming language: it parses human-readable code to machine-readable code, executes it, returns the answer.
- Jupyter is a protocol for interacting with a programming language.
- Jupyter stores inputs and outputs as .ipynb files.
- Jupyter notebooks display inputs and outputs in a browser.
- JuliaBox is an interface to a webserver running Julia.
Julia and Jupyter

- Julia is a programming language: it parses human-readable code to machine-readable code, executes it, returns the answer.
- Jupyter is a protocol for interacting with a programming language.
- Jupyter stores inputs and outputs as .ipynb files.
- Jupyter notebooks display inputs and outputs in a browser.
- JuliaBox is an interface to a webserver running Julia.

how to access?

- recommended: install anaconda distribution, then julia (go to section or see section materials for details)
- not recommended: use Google Colab
  https://discourse.julialang.org/t/julia-on-google-colab-free-gpu-accelerated-shareable-notebooks/15319
Summary statistics

univariate

▶ mean, median, mode
▶ max, min, range
▶ variance
▶ ...

explore via Julia + Jupyter notebook

https://github.com/ORIE4741/demos/blob/master/eda.ipynb
Summary statistics

univariate

- mean, median, mode
- max, min, range
- variance
- ...

explore via Julia + Jupyter notebook

https://github.com/ORIE4741/demos/blob/master/eda.ipynb

multi- (but usually just bi-)variate

- correlation, covariance
- ...

# The perils of summary statistics

Here is a table showing the same mean, variance, correlation, line of best fit...

<table>
<thead>
<tr>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>y</td>
<td>x</td>
<td>y</td>
</tr>
<tr>
<td>10.0</td>
<td>8.04</td>
<td>10.0</td>
<td>9.14</td>
</tr>
<tr>
<td>8.0</td>
<td>6.95</td>
<td>8.0</td>
<td>8.14</td>
</tr>
<tr>
<td>13.0</td>
<td>7.58</td>
<td>13.0</td>
<td>8.74</td>
</tr>
<tr>
<td>9.0</td>
<td>8.81</td>
<td>9.0</td>
<td>8.77</td>
</tr>
<tr>
<td>11.0</td>
<td>8.33</td>
<td>11.0</td>
<td>9.26</td>
</tr>
<tr>
<td>14.0</td>
<td>9.96</td>
<td>14.0</td>
<td>8.10</td>
</tr>
<tr>
<td>6.0</td>
<td>7.24</td>
<td>6.0</td>
<td>6.13</td>
</tr>
<tr>
<td>4.0</td>
<td>4.26</td>
<td>4.0</td>
<td>3.10</td>
</tr>
<tr>
<td>12.0</td>
<td>10.84</td>
<td>12.0</td>
<td>9.13</td>
</tr>
<tr>
<td>7.0</td>
<td>4.82</td>
<td>7.0</td>
<td>7.26</td>
</tr>
<tr>
<td>5.0</td>
<td>5.68</td>
<td>5.0</td>
<td>4.74</td>
</tr>
</tbody>
</table>
The perils of summary statistics

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>y</td>
<td>x</td>
<td>y</td>
<td>x</td>
</tr>
<tr>
<td></td>
<td>10.0</td>
<td>8.04</td>
<td>10.0</td>
<td>9.14</td>
</tr>
<tr>
<td></td>
<td>8.0</td>
<td>6.95</td>
<td>8.0</td>
<td>8.14</td>
</tr>
<tr>
<td></td>
<td>13.0</td>
<td>7.58</td>
<td>13.0</td>
<td>8.74</td>
</tr>
<tr>
<td></td>
<td>9.0</td>
<td>8.81</td>
<td>9.0</td>
<td>8.77</td>
</tr>
<tr>
<td></td>
<td>11.0</td>
<td>8.33</td>
<td>11.0</td>
<td>9.26</td>
</tr>
<tr>
<td></td>
<td>14.0</td>
<td>9.96</td>
<td>14.0</td>
<td>8.10</td>
</tr>
<tr>
<td></td>
<td>6.0</td>
<td>7.24</td>
<td>6.0</td>
<td>6.13</td>
</tr>
<tr>
<td></td>
<td>4.0</td>
<td>4.26</td>
<td>4.0</td>
<td>3.10</td>
</tr>
<tr>
<td></td>
<td>12.0</td>
<td>10.84</td>
<td>12.0</td>
<td>9.13</td>
</tr>
<tr>
<td></td>
<td>7.0</td>
<td>4.82</td>
<td>7.0</td>
<td>7.26</td>
</tr>
<tr>
<td></td>
<td>5.0</td>
<td>5.68</td>
<td>5.0</td>
<td>4.74</td>
</tr>
</tbody>
</table>

same mean, variance, correlation, line of best fit...
The perils of summary statistics
The perils of summary statistics: modern update

https://www.autodeskresearch.com/publications/samestats
What to visualize?

- examples across all features (usually not)
- plot features across all examples (much more common)
Best practices

- Always label your axes.
- Ensure all marks on plot are meaningful.
- Beware of pie charts; bar charts are often easier to read.
- Beware of line plots; if your data is not continuous, try scatter plot instead.
- Consider which curves to plot on same axes. Make comparisons easy!
Take away

- always look at (some of) your data
- decide what question you want to answer
Questions?

https://docs.google.com/spreadsheets/d/1vLbwi0WCOo0wU6cU_r0RHA_nY7C0fDZ1F8Yq09pqYYuk